REMARKS

Reconsideration and allowance in view of the foregoing amendment and the following remarks are respectfully requested.

Rejection of Claims 22-25, 27, 29-32 and 34 Under 35 U.S.C. §103(a)

The Office Action rejects claims 22-25, 27, 29-32 and 34 under 35 U.S.C. §103(a) as being unpatentable over Ezzat et al. (Visual Speech Synthesis by Morphing Visemes) ("Ezzat et al.") in view of Jiang et al. (Visual Speech Analysis with Application to Mandarin Speech Training) ("Jiang et al.") in view of Hon et al. (Automatic Generation of Synthesis Units for Trainable Text-to-Speech Systems) ("Hon et al."). Applicants thank the Examiner for the detailed discussion in the Advisory Action. Applicants maintain their position that under an appropriate obviousness analysis, that one of skill in the art would not have a sufficient amount of motivation (by a preponderance of the evidence) to modify the teachings of Ezzat et al. with the teachings of Hon et al.

Applicants first address some of the particular arguments in the Advisory Action.

First, Advisory Action states in the discussion of Ezzat et al. and Hon et al. that "given that both of these references deal with generating audio as well as video streams (as explained in previously written office actions), the Examiner does not see how one of skill in the art can find these references to be in such conflict as to suggest teachings away from one another." (Emphasis added.) Applicants respectfully correct the foundation of the Examiner's conclusion that he cannot see how one of skill in the art could find conflict in these teachings. This paragraph asserts that previous written office actions have established that Hon et al. teach generating audio "as well as video streams". Applicants would surmise that the Examiner perhaps is mistakening the teachings of Hon et al. with Jiang et al. which do focus on an aspect of visual speech analysis and synthesis. However, Applicants strongly submit that Hon et al. fail

to teach anything dealing with video streams and exclusively focuses on audio in audio synthesis and in the Whistler Text-to-Speech engine. Accordingly, the Examiner cannot rely on this erroneous interpretation of Hon et al. as a foundation to combine Hon et al. with Ezzat et al. Applicants suppose that with a correct interpretation of the teachings of Hon et al., that this certainly opens the way for the Examiner to conceive of why the suggestive power of each reference can teach away from their combination. Applicants shall provide further details on why this issue is important in terms of whether it is obvious to combine these references inasmuch as the fact that Ezzat et al. utilize visual speech synthesis and Hon et al. fail to do so.

Furthermore, inasmuch as the Advisory Action uses an erroneous characterization of Hon et al. as a foundation of the entire argument that these references are "similar in their scope", Applicants respectfully request either a Notice of Allowance based on previous arguments and arguments set forth herein or a non-final Office Action and a return of our fee for filing this RCE.

The Advisory Action also characterizes in several places the teachings of Hon et al. incorrectly. For example, the second sentence of the last paragraph states "However, in the context of Hon, the criticism of Hon with diphones is referring to is [sic] the selection or decision process used to match diphones in existing diphone systems", citing the first paragraph under Section 2.1. We have discussed this paragraph at length and Applicants note that this paragraph of Hon et al. does not simply refer to the "selection or decision process" that is used to match diphones, but discusses the use of the diphone as a synthesis unit for concatenative synthesizers. Accordingly, the correct interpretation of this paragraph is not that it is referring to a "process used to match diphones", but rather the difficulty in using diphones as the synthesis unit in a TTS system. For example, this paragraph of Hon et al. teaches "while diphones retain the transitional

information, there can be large distortions due to the difference in spectra between the stationary parts of two units obtained from different context." Several examples are then given.

Similarly, several sentences into the last paragraph of the Advisory Action state a similar characterization of the teachings of Hon:

> "Thus, Hon is not criticizing the use of phonemes (which are essential [sic] a unit of sound) but rather the process to select or match them through diphones (which is pair or transition of phonemes)."

Again, it appears in this language of the Advisory Action that the Examiner misconstrues the reference to diphones in Hon et al. Again, the large distortions due to the different spectra between the stationary parts of two units obtained through different context as the identified problem with using diphones as the synthesis unit in Hon et al. does not relate to the process of selecting or matching phonemes in a database. Applicants certainly note that Hon et al. does not criticize the use of phonemes in general, but rather criticizes the use of diphones as the synthesis unit in a unit selection process. This is the very reason (in connection with the fact the Hon et al. fails to teach anything regarding visual synthesis) that one of skill in the art would less likely to modify the teachings of Ezzat et al. with Hon et al. because Hon et al. highlight the benefits of using diphones as their synthesis unit.

Next, the Advisory Action recites some the arguments in previous Office Actions to articulate the enhancements in the technology of Ezzat et al., which the Examiner asserts would be enhanced by the teachings of Hon et al. In several places, the Examiner references page 4 of the previous Office Action mailed out 6/26/2007. In this portion of the Final Office Action, the Examiner asserts "Ezzat does not teach the claimed 'unit selection process' and does not teach the claimed 'in which a longest possible candidate image sample is selected'". The Office Action then asserts that Hon et al. teach a unit selection process by teaching of "unit selection"

and suggests the claimed limitation of the longest possible candidate image sample being selected.

Applicants respectfully traverse this analysis and note that the basic concept of a unit selection process in text-to-speech is merely the selection of units of speech that are then concatenated together to produce the synthesized speech. In this regard, we note that Ezzat et al. teach a text-to-visual speech (TTVS) synthesis system which includes a discussion of concatenating diphones together. See Section 7, first paragraph. Applicants note that while Ezzat et al. does not necessarily expressly teach "a unit selection process" one of skill in the art would already understand the basic process of selecting units (or diphones) and concatenate them together as is done in the Festival TTS system, for example, it is easy to find on the internet references to the Festival TTS system that uses unit selection. Accordingly, the concept asserted in the Office Action that because Ezzat et al. fail to teach "unit selection", one of skill in the art would have motivation to utilize the unit selection features from Hon et al. is erroneous because one of skill in the art would already recognize and be familiar with the Festival TTS system reference by Ezzat et al. and already understand that such basic system already uses unit selection. Accordingly, such a person of skill in the art would not go searching through other references such as Hon et al. for the purpose of enhancing the teachings of Ezzat et al. with a unit selection process. The Advisory Action states that "this improvement (the unit selection approach of Hon et al.)" in the selection process to match phonemes is the very reason why Hon is combined with Ezzat. Inasmuch as Applicants have demonstrated that the reason why the Examiner asserts one of skill in the art would combine Hon et al. with Ezzat et al. is insufficient because Ezzat et al. already inherently has that feature as would be known.

The Advisory Action also states "further, the motivation provided for combining the prior art is the improvement in the decision or matching process that Hon offers (bottom of page 4 in

Office Action) through their decision tree." On the bottom of page 4 of the final Office Action, the Examiner states that the advantage "to the combination is that with Hon, unit selection features selected from a database of a large amount of candidates can produce optimal concatenation quality" as is mentioned in the first partial paragraph of page 296 of Hon et al.

Applicants again strongly traverse the conclusion that Hon et al. would represent an enhancement over the technology of Ezzat et al. such that one of skill in the art would be motivated to combine these references. Applicants strongly note that the Examiner has failed to analyze Applicants core argument. This argument, as was referenced briefly above, relates to the fact that only Ezzat et al. relate to audio as well as video streams in their visual speech synthesis approach. As Applicants have previously discussed, Ezzat et al. in Section 7 utilize the Festival TTS system that constructs a final audio stream by concatenating diphones together. (First paragraph of Section 7.) As is noted in the last paragraph of Section 7, Ezzat et al. highlight that they "have found the use of TTS timing and phonemic information in *this* manner produces very good quality lip synchronization between the audio and video." (Emphasis added.) The TTS timing and phonetic information is gained through the use of diphones as the selection unit which is "this" manner of selection. Also, at the end of the last paragraph of Section 7, Ezzat et al. note that one of the reasons that diphones are advantageous is that "when a viseme transition is oversampled, the corresponding audio diphone is lengthened to ensure that synchrony between audio and video is maintained." Again, Applicants basic point is that the basic unit that is found to produce "very good" lip synchronization between the audio and the video in Ezzat et al. is the diphone. Thus, while the Examiner has introduced the concept that there are "deficiencies or problems in the selection process" of Ezzat et al., Applicants submit that there is no suggestion within the teachings of Ezzat et al. that using diphones as the unit of selection produces any sort

of deficiency. The suggestion from Ezzat et al. is in fact the opposite, that using diphones as the selection unit is optimal when synchronizing audio with video.

Furthermore, Applicants again submit that it would certainly not be obvious to incorporate the teachings of the selection process of Hon et al. that abandons the diphone as the basic selection unit and requires the use of a decision tree cluster phone-based unit which may be one of a triphone, quinphone, stress-sensitive phone, word-dependent phone or a combination of the above. (See first paragraph, Section 2.3) One reason why one of skill in the art would not be likely or motivated to incorporate this particular process into the visual speech synthesis process of Ezzat et al. is that there may be a myriad of challenges that would be introduced into a visual synthesis approach when it comes to maintaining "very good quality lip synchronization between the audio and the video." For example, how would one of skill in the art have to modify the opportunity to deal with the situation identified at the end of Section 7 of Ezzat et al. related to when a visual viseme transition is oversampled? Is it possible to take a corresponding audio triphone, quinphone, stress-sensitive phone, or word-dependent phone and simply lengthen it to maintain a synchronous relationship between the audio and video? Such a requirement would certainly require further research and thought and is not simply cannot be a matter of replacing diphones as selection units with a completely different decision tree clustered phone-based unit. Inasmuch as Hon et al. fail to teach anything regarding visual synthesis, there is nothing in Hon et al. that would suggestion that its approach would provide a benefit to the teachings of Ezzat et al. and may more likely introduce synchronization problems into Ezzat et al.

Applicants therefore respectfully submit that by a preponderance of the evidence Applicants have the weightier arguments against there being sufficient motivation or suggestion to combine these references. Applicants have corrected foundational assertions within the Advisory Action that cause the conclusions to become much less persuasive inasmuch as they

are based on a technically incorrect characterization of the teachings of either Hon et al. or Ezzat et al. (i.e., Hon et al. does not deal with audio "as well as video streams" and Hon et al. does not identify any deficiencies or problems in their diphone selection process, but rather highlight it as producing very good quality lip synchronization).

Accordingly, Applicants respectfully submit that the preponderance of the evidence is against the combination of these references. Therefore, Applicants request a Notice of Allowance.

CONCLUSION

Having addressed all rejections and objections, Applicants respectfully submit that the subject application is in condition for allowance and a Notice to that effect is earnestly solicited. If necessary, the Commissioner for Patents is authorized to charge or credit the **Novak, Druce & Quigg, LLP, Account No. 14-1437** for any deficiency or overpayment.

Respectfully submitted,

Date: September 25, 2007

By: _____

Thomas M. Isaacson

Attorney for Applicants
Reg. No. 44,166
Phone: 410-286-9405
Fax No.: 410-510-1433

Correspondence Address:
Thomas A. Restaino
Reg. No. 33,444
AT&T Corp.
Room 2A-207
One AT&T Way
Bedminster, NJ 07921